sending performance data of a first server to a second server, said first server being part of said cluster of servers;

based on the performance data, said second server determining if said first server has reached a predetermined upper level of utilization; and

if said first server has reached said predetermined upper level of utilization, said second server automatically sending a reconfiguration request to a server responsible for allocating servers to said cluster to allocate another server to said cluster, and in response, said responsible server automatically identifying another, available server and connection information for said other server and allocating said other server to said cluster.

23.(New) A method as set forth in claim 22 wherein the step of automatically allocating said other server to said cluster comprises the steps of updating a configuration file of said responsible server to list said other server as part of said cluster.

24. (New) A method as set forth in claim 22 further comprising the steps of:

based on the performance data, said second server determining if said first server is under utilized; and

if said first server is under utilized, said second server automatically sending a reconfiguration request to said server responsible for allocating servers to said cluster to de-allocate said first server from said cluster, and in response, said responsible server automatically de-allocating said first server from said cluster.

25. (New) A method as set forth in claim 24 wherein the step of automatically de-allocating said first server from said cluster comprises the step of updating a configuration file of said responsible server to remove said first server from said cluster.

26. (New) A system for allocating servers to a cluster of servers, said system comprising:

means for sending performance data of a first server to a second server, said first server being part of said cluster of servers;

means, based on the performance data, within said second server for determining if said first server has reached a predetermined upper level of utilization, and if said first server has reached said predetermined upper level of utilization, automatically sending a reconfiguration request to a server responsible for allocating servers to said cluster to allocate another server to said cluster; and

means, responsive to said reconfiguration request, within said responsible server for automatically identifying another, available server and connection information for said other server and allocating said other server to said cluster.

27. (New) A system as set forth in claim 26 further comprising:

means, based on the performance data, within said second server for determining if said first server is under utilized, and if said first server is under utilized, automatically sending a reconfiguration request to said server responsible for allocating servers to said cluster to de-allocate said first server from said cluster; and

means, responsive to the de-allocation reconfiguration request, within said responsible server for automatically de-allocating said first server from said cluster.

28. (New) A method for managing servers, said method comprising the steps of:

a first server determining performance data for said first server and performance data for a second server, and reporting to a third server said performance data for said first server and said

performance data for said second servers, said first and second servers being in a cluster of servers;

based on the reported performance data, said third server determining if said first server or said second server has reached a predetermined upper level of utilization; and

if said first server or said third server has reached said predetermined upper level of utilization, said third server sending a reconfiguration request to said first server to reduce subsequent utilization of the server which has reached said predetermined upper level of utilization, and said first server automatically reconfigurating itself to reduce subsequent utilization of the server which has reached said predetermined upper level of utilization.

29. (New) A method as set forth in claim 28 further comprising the earlier steps of:

said first server receiving a request from a client device and determining whether said first server should handle said request, and if so, said first server handling said request, and if not, said first server identifying said second server as available to handle said request and forwarding the request to said second server for handling.

30. (New) A method as set forth in claim 28 wherein the step of said first server automatically reconfigurating itself comprises the step of said first server updating a configuration file which lists one or more servers which are available to handle specified types of client requests.

31. (New) A method as set forth in claim 28 wherein the step of said first server automatically reconfigurating itself comprises the step of automatically identifying a fourth, available server and connection information for said fourth server and allocating said fourth server to said cluster.

32. (New) A method as set forth in claim 28 further comprising the steps of:

based on said performance data, said third server determining if said first server or said second server is under utilized; and

if said first server or said third server is under utilized, said third server sending a reconfiguration request to said first server to de-allocate the server which is under utilized, and said first server automatically reconfiguring itself to de-allocate from said cluster said server which is under utilized.

33. (New) A method as set forth in claim 32 wherein the automatic de-allocating reconfiguring step comprises the step of adding to a pool of available servers, said server which is under utilized, such that if a server in said cluster subsequently reaches said predetermined upper level of utilization, the de-allocated server can be re-allocated to said cluster.

34. (New) A method as set forth in claim 28 wherein:

said first server reports to said third server said performance data for said first server and said performance data for said second server using XML data streams; and

said third server sends said reconfiguration request to said first server using an XML data stream.

35. (New) A method as set forth in claim 29 wherein the step of said first server determining whether said first server should handle said request comprises the step of said first server determining if said request is for data which is currently cached at said first server, and if so, the step of said first server handling said request comprises the step of said first server supplying said data from the cache at said first server to said client device.--